A. Optimal transportation : discrete case

A.1. Optimal transport

When considering a limited number of samples for the two distribution, the computation of the Wasserstein distance can be solved through linear programming algorithms. In the balanced case, we have $X = \{x_1, \dots, x_{2n}\}$ where $\{x_1, \dots, x_n\}$ are sampled from P_+ and $\{x_{n+1}, \dots, x_{2n}\}$ are sampled from P_-. We note $U = \{u_1, \dots, u_{2n}\}$ the labels with $u_1, \dots, u_n = 1$ and $u_{n+1}, \dots, u_{2n} = -1$ and C the $n \times n$ matrix cost function with $C_{i,j} = ||x_i - x_{n+j}||$. The primal problem of the optimal transport is to find a transportation plan Π (a $n \times n$ matrix) such as:

$$\min_{\Pi} \quad \sum_{i,j \in n \times n} \Pi_{i,j} * C_{i,j} \tag{13}$$

$$\text{subject to} \quad \Pi_{i,j} \geq 0, \tag{14}$$

$$\sum_i \Pi_{i,j} = \frac{1}{n}, \sum_j \Pi_{i,j} = \frac{1}{n}. \tag{15}$$

The constraints enforce Π to be a discrete joint probability distribution with the appropriate marginals as in the continuous case. The dual formulation for the discrete optimal transport problem is:

$$\max_{F} \quad F.U^T \tag{16}$$

$$\text{subject to} \quad \forall i, j \in n \times n, F_i - F_{n+j} \leq C_{i,j} \tag{17}$$

where F is a $2n$ vector that is a discrete version of the function f of Equation 1b. The constraint on F is the discrete counterpart of the 1-Lipschitz constraint.

A.2. Hinge regularized Optimal transport

Similarly to the classical case, the discrete counterpart of the regularized Wasserstein distance is also a transportation problem which has the following formulation:

$$\begin{aligned}
\min_{\Pi} \quad & \sum_{i,j \in n \times n} [\Pi_{i,j} * C_{i,j}] - 2\left(1 - \sum_{i,j \in n \times n} [\Pi_{i,j}]\right) \\
\text{subject to} \quad & \Pi_{i,j} \geq 0, \\
& \frac{1}{n} \leq \sum_i \Pi_{i,j} \leq \frac{1+\lambda}{n}, \\
& \frac{1}{n} \leq \sum_j \Pi_{i,j} \leq \frac{1+\lambda}{n}.
\end{aligned}$$

Roughly speaking, it allows to give more weight to the transportation of the closest pairs by admitting to deviate from the marginals with a tolerance that depends on λ. Since the closest pairs in the two distributions are the most difficult to classify, it illustrates why this formulation is more adequate for classification problems. The dual formulation of this transportation problem is a discrete counterpart of Equation 5 :

$$\begin{aligned}
\max_{F} \quad & \sum_{k=0}^{2n} [F_i * u_i - \lambda(0, 1 - F_i * u_i)_+] \\
\text{subject to} \quad & \forall i, j \in n \times n, F_i - F_{n+j} \leq C_{i,j}.
\end{aligned}$$

We observe that the constraint in the dual problem are not affected by the new formulation and still corresponds to a the 1-Lipschitz constraint in the continuous case.

B. Theorem proofs

B.1. Proof Theorem 1

We denote as

$$f^* := f^*_\lambda \in \arg\min_{f\in \mathrm{Lip}_1(\Omega)} \mathcal{L}^{hKR}_\lambda(f) \quad \text{and} \quad \hat{f}_n := \hat{f}_{n,\lambda} \in \arg\min_{f\in \mathrm{Lip}_1(\Omega)} \hat{\mathcal{L}}^{hKR}_{\lambda,n}(f). \tag{18}$$

If we assume that (6) is not true, then there exists $\mathbf{x} \in \Omega$ such that $f^*(\mathbf{x}) > 1 + \mathrm{diam}(\Omega) + \frac{\mathcal{R}(\psi)}{\inf(p,1-p)}$ or $f^*(\mathbf{x}) < -1 - \mathrm{diam}(\Omega) - \frac{L_1(\psi)}{\inf(p,1-p)}$. We suppose without loss of generality that the first inequality holds. If $\mathbf{z} \in \Omega$ then the 1-Lipschitz condition in f^* implies that $f^*(\mathbf{z}) > 1 + \frac{L_1(\psi)}{1-p}$. Hence $(1 - f^*)_+ = 0$ and

$$\begin{aligned}
L(f^*) &\leq \sup_{g\in Lip_1(\Omega)} L_2(g) - \lambda L_1(f^*) \\
&= \sup_{g\in Lip_1(\Omega)} E_{X|Y=1}(g(X)) - E_{X|Y=-1}(g(X)) - E\left\{\lambda(1 - Y f^*(X))_+\right\} \\
&= L_2(\psi) - \lambda\{pE_{X|Y=1}(1 - f^*(X))_+ + (1-p)E_{X|Y=-1}(1 + f^*(X))_+\} \\
&\leq L_2(\psi) - \lambda\{(1-p)E_{X|Y=-1}(1 + f^*(X))\} \\
&\leq L_2(\psi) - \lambda\{(1-p)E_{X|Y=-1}(2 + \frac{L_1(\psi)}{1-p})\} \\
&= L_2(\psi) - 2\lambda(1-p) - \lambda\{E_{X|Y=-1}(L_1(\psi))\} = L_2(\psi) - 2\lambda(1-p) - \lambda L_1(\psi)
\end{aligned}$$

Then f^* can not be an optimal solution of the problem (18). Then there exists some constant M large enough, such that f^* belongs to $\mathrm{Lip}_1^M(\Omega) := \{f \in \mathrm{Lip}_1(\Omega) : ||f||_\infty \leq M\}$ and not to $\mathrm{Lip}_1(\Omega)$. Since the functional $\mathcal{L}^{hKR}_\lambda$ is convex and $\mathrm{Lip}_1^M(\Omega)$ is compact in $\mathcal{C}(\Omega)$, we are able to make use of Ascoli-Arzela Theorem and conclude that there exists at least one function minimizing the expected loss. Furthermore the set of those functions is compact and convex.

B.2. Proof Theorem 2

Definition B.1. *Let μ, ν two positive measures in $\mathbb{R}^d$. The Kullback-Leibler divergence from μ to ν is defined as*

$$KL(\mu|\nu) = \begin{cases} \int \log(\frac{d\mu}{d\nu})d\mu - \int d\mu + \int d\nu & \text{if } \mu << \nu \\ \infty & \text{otherwise} \end{cases} \tag{19}$$

Theorem 3. *Let $\phi_1, \phi_2 : \Omega \to \bar{\mathbb{R}}$ be lower semicontinuous convex functions and $\mu, \nu \in \mathcal{P}(\Omega)$ be probability measures. Then for all $\epsilon > 0$ the following equality holds*

$$\begin{aligned}
&\inf_{\pi\in\Pi_+(\mu,\nu)} \int \phi_1(-\frac{d\pi_{\boldsymbol{x}}}{d\mu}(\boldsymbol{x}))d\mu(\boldsymbol{x}) + \int \phi_2(-\frac{d\pi_{\boldsymbol{z}}}{d\nu}(\boldsymbol{z}))d\nu(\boldsymbol{z}) + \epsilon KL(\pi|e^{-\frac{c(\boldsymbol{x},\boldsymbol{z})}{\epsilon}}(d\mu \times d\nu)) \\
&= \sup_{f,g\in L^1(\Omega)} -\int_\Omega \phi_1^*(f(\boldsymbol{x}))d\mu(\boldsymbol{x}) - \int_\Omega \phi_1^*(g(\boldsymbol{z}))d\nu(\boldsymbol{z}) - \epsilon \int \left(e^{\frac{f(\boldsymbol{x})+g(\boldsymbol{z})-c(\boldsymbol{x},\boldsymbol{z})}{\epsilon}} - e^{-\frac{c(\boldsymbol{x},\boldsymbol{z})}{\epsilon}}\right) d\mu d\nu.
\end{aligned} \tag{20}$$

Furthermore if $\epsilon = 0$ then

$$\begin{aligned}
&\inf_{\pi\in\Pi_+(\mu,\nu)} \int_{\Omega\times\Omega} c(\boldsymbol{x},\boldsymbol{z})d\pi(\boldsymbol{x},\boldsymbol{z}) + \int \phi_1(-\frac{d\pi_{\boldsymbol{x}}}{d\mu}(\boldsymbol{x}))d\mu(\boldsymbol{x}) + \int \phi_2(-\frac{d\pi_{\boldsymbol{z}}}{d\nu}(\boldsymbol{z}))d\nu(\boldsymbol{z}) \\
&= \sup_{(f,g)\in\Phi(\mu,\nu)} -\int_\Omega \phi_1^*(f(\boldsymbol{x}))d\mu(\boldsymbol{x}) - \int_\Omega \phi_1^*(g(\boldsymbol{z}))d\nu(\boldsymbol{z}).
\end{aligned} \tag{21}$$

Where $\Pi_+(\mu,\nu)$ is the set of positive measures $\pi \in \mathcal{M}_+(\Omega \times \Omega)$ which are absolutely continuous with respect to the joint measure $d\mu \times d\nu$, and $\Phi(\mu,\nu)$ consists of the pairs of functions $(f,g) \in L_1(\Omega) \times L_1(\Omega)$ such that $c(\boldsymbol{x},\boldsymbol{z}) - f(\boldsymbol{x}) - g(\boldsymbol{z}) \geq 0$ $d\mu \times d\nu - a.s.$.

First we recall the Fenchel–Rockafellar Duality result, we use a weaker version given in Theorem 1.12 in [6]

Proposition 3. *Let E be a Banach space and $\Upsilon, \Psi : E \to \mathbb{R} \cup \{\infty\}$ be two convex functions, assume that there exist $z_0 \in dom(\Psi) \cap dom(\Upsilon)$ such that Ψ is continous in z_0. Then strong duality holds*

$$\inf_{a \in E} \{\Upsilon(a) + \Psi(a)\} = \sup_{b \in E^*} \{-\Upsilon^*(-b) - \Psi^*(b)\} \tag{22}$$

We identify the different elements of our problem with such of previous Proposition.

- E is the space of continuous functions in $\Omega \times \Omega$. Note that the set is bounded, hence E^*, by Riesz theorem, is the set of regular measures in $\Omega \times \Omega$.
- If $\epsilon = 0$:

$$\Psi_0(u) = \begin{cases} 0 & \text{if } \ u(\mathbf{x}, \mathbf{z}) \geq -c(\mathbf{x}, \mathbf{z}) \\ \infty & \text{otherwise} \end{cases} \tag{23}$$

$$\Upsilon_0(u) = \begin{cases} \int \phi_1^*(-f(\mathbf{x}))d\mu(\mathbf{x}) + \int \phi_2^*(-g(\mathbf{z}))d\nu(\mathbf{z}) & \text{if } \ u(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) \\ \infty & \text{otherwise} \end{cases} \tag{24}$$

If $\epsilon > 0$:

$$\Psi_\epsilon(u) = \epsilon \int \left(e^{\frac{u(\mathbf{x},\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) \tag{25}$$

$$\Upsilon_\epsilon(u) = \Upsilon_0(u) \tag{26}$$

Note that $\Upsilon_\epsilon(u) = \Upsilon_0(u)$ could be non well defined, to avoid this situation we fix $\mathbf{x}_0 \in \Omega$ and consider $u(\mathbf{x}, \mathbf{z}) = (u(\mathbf{x}, \mathbf{z}_0) - u(\mathbf{z}_0, \mathbf{z}_0)/2) + u(\mathbf{z}_0, \mathbf{z}) - u(\mathbf{z}_0, \mathbf{z}_0)/2)$. Now we compute the dual operators

$$\begin{aligned} \Psi_\epsilon^*(-\pi) &= \sup_{u \in E} \left\{ -\epsilon \int \left(e^{\frac{u(\mathbf{x},\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) - \int u(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) \right\} \\ &= \sup_{u \in E} \left\{ -\epsilon \int \left(e^{\frac{u(\mathbf{x},\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) + \int u(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) \right\} \end{aligned}$$

Now if π were not absolutely continuous respect the joint measure $e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} d\mu \times d\nu$ then we would have a continuous function $u(\mathbf{x}, \mathbf{z}) = 0$ $d\mu \times d\nu$ almost surely and such that $\int u(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) \neq 0$. If we take the function $\lambda u(\mathbf{z}, \mathbf{z})$ and λ tends to $\pm\infty$ we deduce that the supremum is ∞. Then suppose that $d\pi = m(\mathbf{x}, \mathbf{z}) e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} (d\mu \times d\nu)$.

$$\begin{aligned} \Psi_\epsilon^*(-\pi) &= \begin{cases} \sup_{u \in E} \left\{ \epsilon \int \left(-e^{\frac{u(\mathbf{x},\mathbf{z})}{\epsilon}} + 1 + \frac{u(\mathbf{x},\mathbf{z})}{\epsilon} m(\mathbf{x}, \mathbf{z}) \right) e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} d\mu(\mathbf{x}) d\nu(\mathbf{z}) \right\} & \text{if } d\pi = m e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} (d\mu \times d\nu). \\ \infty & \text{otherwise.} \end{cases} \\ &= \epsilon KL(\pi | e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} (d\mu \times d\nu)) \end{aligned}$$

With some similar calculation, we compute for $\epsilon = 0$:

$$\Psi_0^*(-\pi) = \begin{cases} \int c(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) & \text{if } \pi \text{ is a positive measure.} \\ \infty & \text{otherwise.} \end{cases}$$

Finally for $\Upsilon_\epsilon^* = \Upsilon_0^*$

$$\begin{aligned} \Upsilon_\epsilon^*(\pi) &= \sup_{u \in E, \ u(\mathbf{x},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z})} \left\{ \int f(\mathbf{x}) + g(\mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) - \int \phi_1^*(-f(\mathbf{x}))d\mu(\mathbf{x}) - \int \phi_2^*(-g(\mathbf{z}))d\nu(\mathbf{z}) \right\} \\ &= \sup_{u \in E, \ u(\mathbf{x},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z})} \left\{ \int f(\mathbf{x}) d\pi_{\mathbf{x}}(\mathbf{x}) - \int \phi_1^*(-f(\mathbf{x}))d\mu(\mathbf{x}) + \int g(\mathbf{z}) d\pi_{\mathbf{z}}(\mathbf{z}) - \int \phi_2^*(-g(\mathbf{z}))d\nu(\mathbf{z}) \right\} \\ &= \sup_{f \in C(\Omega)} \left\{ \int f(\mathbf{x}) d\pi_{\mathbf{x}}(\mathbf{x}) - \int \phi_1^*(-f(\mathbf{x}))d\mu(\mathbf{x}) \right\} + \sup_{g \in C(\Omega)} \left\{ \int g(\mathbf{z}) d\pi_{\mathbf{z}}(\mathbf{z}) - \int \phi_2^*(-g(\mathbf{z}))d\nu(\mathbf{z}) \right\} \\ &= (I_1) + (I_2). \end{aligned}$$

We first consider (I_1). The same reasoning will hold for (I_2). If $\pi_{\mathbf{x}}$ were not absolutely continuous respect μ then reasoning as before we obtain ∞. Then $d\pi_{\mathbf{x}} = \frac{d\pi_{\mathbf{x}}}{d\mu} d\mu$ and

$$
\begin{aligned}
(I_1) &= \sup_{f \in C(\Omega)} \left\{ \int \left(-f(\mathbf{x}) \frac{d\pi_{\mathbf{x}}}{d\mu} - \phi_1^*(f(\mathbf{x})) \right) d\mu(\mathbf{x}) \right\} \\
&= \int \left(\sup_m \left\{ -\frac{d\pi_{\mathbf{x}}}{d\mu} m - \phi_1^*(m) \right\} \right) d\mu(\mathbf{x}) = \int \phi_1(-\frac{d\pi_{\mathbf{x}}}{d\mu}) d\mu(\mathbf{x}) \\
(I_2) &= \int \phi_2(-\frac{d\pi_{\mathbf{x}}}{d\nu}) d\mu(\mathbf{z})
\end{aligned}
$$

Note that the inversion of the supremum and the integral is guaranteed here since $(\mathbf{x}, m) \mapsto -m \frac{d\pi_{\mathbf{x}}}{d\mu}(\mathbf{x}) + \phi_1^*(m)$ is lower semi-continuous and convex in m and measurable in $(\mathbf{x}, m)$. Then it is a normal integrand, and we can apply Theorem 14.60 in [30].
Then computing both in Equation (22) we end with the following result

$$
\begin{aligned}
&\inf_{u(\mathbf{x},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) \geq -c(\mathbf{x},\mathbf{z})} \left\{ \int \phi_1^*(-f(\mathbf{x})) d\mu(\mathbf{x}) + \int \phi_2^*(-g(\mathbf{z})) d\nu(\mathbf{z}) \right\} \\
&\qquad = \inf_{f(\mathbf{x}) + g(\mathbf{z}) \leq c(\mathbf{x},\mathbf{z})} \left\{ \int \phi_1^*(f(\mathbf{x})) d\mu(\mathbf{x}) + \int \phi_2^*(g(\mathbf{z})) d\nu(\mathbf{z}) \right\} \\
&\qquad = - \sup_{f(\mathbf{x}) + g(\mathbf{z}) \leq c(\mathbf{x},\mathbf{z})} \left\{ - \int \phi_1^*(f(\mathbf{x})) d\mu(\mathbf{x}) - \int \phi_2^*(g(\mathbf{z})) d\nu(\mathbf{z}) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\sup_{\pi \in \mathcal{M}_+(\Omega \times \Omega)} \left\{ - \int c(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) - \int \phi_2(-\frac{d\pi_{\mathbf{x}}}{d\nu}) d\mu(\mathbf{z}) - \int \phi_1(-\frac{d\pi_{\mathbf{x}}}{d\mu}) d\mu(\mathbf{x}) \right\} \\
&\qquad = - \inf_{\pi \in \mathcal{M}_+(\Omega \times \Omega)} \left\{ \epsilon \int c(\mathbf{x}, \mathbf{z}) d\pi(\mathbf{x}, \mathbf{z}) + \int \phi_2(-\frac{d\pi_{\mathbf{x}}}{d\nu}) d\mu(\mathbf{z}) + \int \phi_1(-\frac{d\pi_{\mathbf{x}}}{d\mu}) d\mu(\mathbf{x}) \right\}.
\end{aligned}
$$

$$
\begin{aligned}
&\inf_{u(\mathbf{x},\mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z})} \left\{ \epsilon \int \left(e^{\frac{-f(\mathbf{x}) - g(\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) + \int \phi_1^*(-f(\mathbf{x})) d\mu(\mathbf{x}) + \int \phi_2^*(-g(\mathbf{z})) d\nu(\mathbf{z}) \right\} \\
&\qquad = \inf_{f,g} \left\{ \epsilon \int \left(e^{\frac{f(\mathbf{x}) + g(\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) + \int \phi_1^*(f(\mathbf{x})) d\mu(\mathbf{x}) + \int \phi_2^*(g(\mathbf{z})) d\nu(\mathbf{z}) \right\} \\
&\qquad = - \sup_{f,g} \left\{ -\epsilon \int \left(e^{\frac{f(\mathbf{x}) + g(\mathbf{z}) - c(\mathbf{x},\mathbf{z})}{\epsilon}} - e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} \right) d\mu(\mathbf{x}) d\nu(\mathbf{z}) - \int \phi_1^*(f(\mathbf{x})) d\mu(\mathbf{x}) - \int \phi_2^*(g(\mathbf{z})) d\nu(\mathbf{z}) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\sup_{\pi \in \mathcal{M}_+(\Omega \times \Omega)} \left\{ -\epsilon KL(\pi | e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} (d\mu \times d\nu)) - \int \phi_2(-\frac{d\pi_{\mathbf{x}}}{d\nu}) d\mu(\mathbf{z}) - \int \phi_1(-\frac{d\pi_{\mathbf{x}}}{d\mu}) d\mu(\mathbf{x}) \right\} \\
&\qquad = - \inf_{\pi \in \mathcal{M}_+(\Omega \times \Omega)} \left\{ \epsilon KL(\pi | e^{-\frac{c(\mathbf{x},\mathbf{z})}{\epsilon}} (d\mu \times d\nu)) + \int \phi_2(-\frac{d\pi_{\mathbf{x}}}{d\nu}) d\mu(\mathbf{z}) + \int \phi_1(-\frac{d\pi_{\mathbf{x}}}{d\mu}) d\mu(\mathbf{x}) \right\}
\end{aligned}
$$

Proof of Theorem 2 With the same notation of Theorem 3, it is enough to consider, $\mu = P_+$ $\nu = P_-$ and

$$\psi_1(s) = \begin{cases} p - s \quad if \quad s \in [p, p + \lambda p] \\ \\ \infty \quad else. \end{cases} \tag{27}$$

$$\psi_2(s) = \begin{cases} 1 - p - s \quad if \quad s \in [1 - p, 1 - p + \lambda(1 - p)] \\ \\ \infty \quad else. \end{cases} \tag{28}$$

Then for each $f \in L_1(d\mu),\ g \in L_1(d\nu)$

$$\begin{aligned} -\psi_1^*(f(\mathbf{x})) &= -\sup_s\{-\psi_1(s) + f(\mathbf{x})s\} = \inf_s\{\psi_1(-s) - f(\mathbf{x})s\} = \inf_s\{\psi_1(s) + f(\mathbf{x})s\} \\ &= \begin{cases} f(\mathbf{x}) \quad if \quad 1 \le f(\mathbf{x}) \\ \\ f(\mathbf{x}) - p\lambda(1 - f(\mathbf{x})) \quad else. \end{cases} \\ &= f(\mathbf{x}) - p\lambda(1 - f(\mathbf{x}))_+ \\ -\psi_2^*(g(\mathbf{z})) &= f(\mathbf{z}) - (1 - p)\lambda(1 - f(\mathbf{z}))_+. \end{aligned}$$

Note that when $\lambda \ge 0$ the functions $r \mapsto h_1(r) := r - p\lambda(1-r)_+$ and $h_2(r) := r - (1-p)\lambda(1-r)_+$ are nondecreasing. Now if we denote as J the right hand side of (20) then

$$J = \sup_{(f,g)\in\Phi(\mu,\nu)} \int_\Omega h_1(f(\mathbf{x}))d\mu(\mathbf{x}) + \int_\Omega h_2(g(\mathbf{z}))d\mu(\mathbf{z}).$$

We denote as f^d the $d-$conjugate of f defined as the function

$$f^d(r) := \inf_{s\in\Omega}\{|r - s| - f(s)\},$$

see for instance in [12] for a suitable definition. It is clear that $f^{dd} \ge f$, and the equality holds if f is a $d-$concave function since it is said that f is $d-$concave if it is the d-conjugate of another function. Hence using the nondecreasing condition of h we get to

$$J = \sup_{f^{dd}, f^d} \int_\Omega h_1(f^{dd}(\mathbf{x}))d\mu(\mathbf{x}) + \int_\Omega h_2(f^d(\mathbf{z}))d\nu(\mathbf{z}).$$

On the other side $f^d(r) = \inf_{s\in\Omega}\{|r - s| - f(s)\}$ is a limit of a sequence of $1-$Lipschitz functions in Ω, hence it belongs to $\mathrm{Lip}_1(\Omega)$. Using the 1-Lipschitz property and taking $r = s$ in the infimum leads to

$$-f^d(r) \le \inf_{s\in\Omega}\{|r - s| - f^d(s)\} \le -f^d(r).$$

This means that $f^{dd} = -f^d(r)$, hence we have that

$$\begin{aligned} J &= \sup_{(-f^d, f^d)} \int_\Omega h_1(f^{dd}(\mathbf{x}))d\mu(\mathbf{x}) + \int_\Omega h_2(f^d(\mathbf{z}))d\mu(\mathbf{z}). \\ &\le \sup_{f\in\mathrm{Lip}_1(\Omega)} \int_\Omega h_1(f^{dd}(\mathbf{x}))d\mu(\mathbf{x}) + \int_\Omega h_2(-f(\mathbf{z}))d\nu(\mathbf{z}) \le J \end{aligned}$$

where the last inequality comes from the fact that if $f \in \mathrm{Lip}_1(\Omega)$ then $(f, -f) \in \Phi(\mu, \nu)$.

B.3. Proof Proposition 1

Even though the proof of Proposition 1 can be done following the frame of the proof of Proposition 1 in [16], we have provided here an easier proof in order to make this document self-content. The proof of this Proposition requires some properties on the transport plan.

Definition B.2. *A set $\Gamma \subset R^d \times \mathbb{R}^d$ is said to be d-cyclically monotone if for all $n \in \mathbb{N}$ and $\{(x_k, y_k)\}_{k=1}^n \subset \Gamma$ it is satisfied*

$$\sum_{k=1}^{n} c(x_k, y_k) \leq \sum_{k=1}^{n} c(x_{k+1}, y_k), \quad \text{assuming that } n+1 = 1. \tag{29}$$

It is said that a measure is $d-$cyclically monotone if its support is $d-$cyclically monotone.

In particular the optimal transference plan in Kantorovich problem for the cost d is $d-$cyclically monotone, see Theorem 2.3 [12]. The same characterization holds for the optimal measures of (20), this claim is proved in the following Lemma.

Lemma 4. *The optimal measure π of (20) is $d-$cyclically monotone for $d(\boldsymbol{x}, \boldsymbol{z}) = ||\boldsymbol{x} - \boldsymbol{z}||$.*

If π were not $d-$cyclically monotone, in [37] it is built another measure $\tilde{\pi}$, with the same marginals as π, such that the value of $\int |\mathbf{x} - \mathbf{z}| d\pi(\mathbf{x}, \mathbf{z}) > \int |\mathbf{x} - \mathbf{z}| \tilde{\pi}(\mathbf{x}, \mathbf{z})$. Computing this we deduce

$$\inf_{\pi \in \Pi_\lambda^p(\mu,\nu)} \int_{\Omega\times\Omega} |\mathbf{x} - \mathbf{z}| d\pi + \pi_{\mathbf{x}}(\Omega) + \pi_{\mathbf{z}}(\Omega) - 1 > \inf_{\tilde{\pi} \in \tilde{\pi}_\lambda^p(\mu,\nu)} \int_{\Omega\times\Omega} |\mathbf{x} - \mathbf{z}| d\tilde{\pi} + \tilde{\pi}_{\mathbf{x}}(\Omega) + \tilde{\pi}_{\mathbf{z}}(\Omega) - 1.$$

Hence π cannot be optimal.
We replicate this construction on order to build this proof as self content as possible.
If P_+ and P_- are discrete probabilities. Then $P_+ = \sum_{k=1}^n u_k \delta_{\mathbf{x}_k}$ and $P_- = \sum_{j=1}^n v_j \delta_{\mathbf{z}_j}$ then the optimal measure has the form:

$$\frac{1}{n} \sum_{k,j=1}^{n} \pi_{k,j} \delta_{\mathbf{x}_k, \mathbf{z}_j} \tag{30}$$

If it is not d-cyclically monotone then there exist $N \in \mathbb{N}$ and $\{(\mathbf{x}_{k_i}, \mathbf{z}_{k_i})\}_{i=1}^N \subset \text{supp}(\pi)$ such that:

$$\sum_{i=1}^{N} ||\mathbf{x}_{k_i} - \mathbf{z}_{k_{i+1}}|| < \sum_{i=1}^{N} ||\mathbf{x}_{k_i} - \mathbf{z}_{k_i}||, \quad \text{assuming that } k_{N+1} = k_1.$$

Let $a := \inf_{i=1,\dots,N}\{\pi_{k_i,k_i}\} > 0$. And let's define $\tilde{\pi}$ as

$$\tilde{\pi} := \pi + \frac{1}{n} \sum_{i=1}^{n} \left(\delta_{\mathbf{x}_{k_i}, \mathbf{z}_{k_{i+1}}} - \delta_{\mathbf{x}_{k_i}, \mathbf{z}_{k_i}} \right).$$

Then

$$\tilde{\pi}(A \times \Omega) = \pi(A \times \Omega) + \frac{1}{n} \sum_{i=1}^{n} \left(\delta_{\mathbf{x}_{k_i}}(A) - \delta_{\mathbf{x}_{k_i}}(A) \right) = \pi(A \times \Omega).$$

And the same holds with $(\Omega \times B)$ and the other marginal, and also it satisfied that

$$\frac{1}{n} \sum_{k,j=1}^{n} |\mathbf{x}_k - \mathbf{z}_j| \tilde{\pi}_{k,j} < \frac{1}{n} \sum_{k,j=1}^{n} ||\mathbf{x}_k - \mathbf{z}_j|| \pi_{k,j}.$$

Hence $\tilde{\pi}$ is the searched measure in the discrete case.
$\Pi_\lambda^p(\mathcal{S}, \mathcal{T})$ is sequentially compact respect the weak convergence denoted * of measures if both $\mathcal{S}$, $\mathcal{T}$ are also. Because of the compactness of $\Omega \times \Omega$, we only have to check that the set is bounded in total variation. But this is

straightforward because for each $\pi \in \Pi_\lambda^p(P_+, P_-)$ it is satisfied $|\pi|(\Omega \times \Omega) \leq (p + p\lambda)(p + p\lambda)$.
If P_+ and P_- are general probabilities. Let $X_1^+, \ldots, X_n^+$ and $Z_1^+, \ldots, Z_n^+$ be sequences of independent random variables with law P_+ and P_-. And let $P_n^+,\ P_-^n$ be the associated empirical measures. Buy using the strong law of large numbers we deduce that $P_n^+ \to P_+$ and $P_-^n \to P_-$ with probability one.
Now let π_n be the corresponding optimal measure for $P_n^+,\ P_-^n$, then there exist a measure π such that $\pi_n \rightharpoonup^* \pi$. It means that for each continuous and bounded function f in $\Omega \times \Omega$ we get

$$\int f d\pi_n \longrightarrow \int f d\pi.$$

Since the norm $(\mathbf{x}, \mathbf{z}) \mapsto ||\mathbf{x} - \mathbf{z}||$ is continuous and bounded, once again because Ω is compact, we derive that

$$\int ||\mathbf{x} - \mathbf{z}|| d\pi_n + \pi_{\mathbf{x}_n}(\Omega) + \pi_{\mathbf{z}_n}(\Omega) - 1 \longrightarrow \int ||\mathbf{x} - \mathbf{z}|| d\pi + \pi_{\mathbf{x}}(\Omega) + \pi_{\mathbf{z}}(\Omega) - 1$$

Finally it is known that if a sequence of measures is d-cyclically monotone and converges weak* to another measure, then it is also d-cyclically monotone. This concludes the proof.

The proof of Proposition 1 is achieved as follows. The assumption of d-cyclically monotone involves that in particular $g(\mathbf{x}) - g(\mathbf{z}) = ||\mathbf{x} - \mathbf{z}||$ π-a.s. for some function g. Then for the balanced case

$$\begin{aligned}
&\int (g-1) d\pi_x - \int (g+1) d\pi_z + 2 \\
&= \sup_{f \in \mathrm{Lip}_1(\Omega)} \int_\Omega f(dP_+ - dP_-) - \lambda \left(\int_\Omega (1-f)_+ dP_+ + \int_\Omega (1+f)_+ dP_- \right).
\end{aligned}$$

Then we split $(g-1) = (g-1)\mathbb{1}_{g-1 \geq 0} + (g-1)\mathbb{1}_{g-1<0}$ and

$$\begin{aligned}
&\int (g-1) d\pi_x + 1 \\
&= (1+\lambda) \int (g-1)\mathbb{1}_{g-1\geq 0} dP_+ + \int (g-1)\mathbb{1}_{g-1<0} dP_+ = \int (g-1) - \lambda(1-g)_+ dP_+.
\end{aligned}$$

Doing the same with P_-, we deduce that this g is optimal and $g(\mathbf{x}) - g(\mathbf{z}) = ||\mathbf{x} - \mathbf{z}||$ π-a.s. for the optimal measure π. As a consequence of such observations, following exactly the same arguments of the proof of Proposition 1 in [16], note that the key is $g(\mathbf{x}) - g(\mathbf{z}) = ||\mathbf{x} - \mathbf{z}||$ π-a.s. which comes from what follows.

Let f^* be the optimal of Lemma 4, $\mathbf{x}$ be a differentiable point of f^*. By assumption, the density property implies that $\pi(\mathbf{x} = \mathbf{z}) = 0$, and then with probability one, there exist $\mathbf{z}$ such that $f^*(\mathbf{x}) - f^*(\mathbf{z}) = ||\mathbf{x} - \mathbf{z}||$ and both points are different $\mathbf{x} \neq \mathbf{z}$. For each $t \in [0,1]$ let $\mathbf{x}_t = (1-t)\mathbf{x} + t\mathbf{z}$ and the path $\sigma : [0,1] \to \mathbb{R}$ defined as $\sigma(t) := f^*(\mathbf{x}_t) - f^*(\mathbf{x})$. The proof is split in two steps;
Step 1 ($\sigma(t) = ||\mathbf{x}_t - \mathbf{z}|| = t||\mathbf{x} - \mathbf{z}||$)
First of all we realize that for each $s, t \in [0,1]$

$$|\sigma(t) - \sigma(s)| = |f^*(\mathbf{x}) - f^*(\mathbf{x}_s)| \leq ||\mathbf{x}_t - \mathbf{x}_s|| \leq |t-s|||\mathbf{x} - \mathbf{z}||.$$

Actually if we consider $t \in [0,1]$ then

$$\begin{aligned}
\sigma(1) - \sigma(0) &\leq \sigma(1) - \sigma(t) + \sigma(t) - \sigma(0) \\
&\leq (1-t)||\mathbf{x} - \mathbf{z}|| + \sigma(t) - \sigma(0) \\
&\leq (1-t)||\mathbf{x} - \mathbf{z}|| + t||\mathbf{x} - \mathbf{z}|| = ||\mathbf{x} - \mathbf{z}|| = \sigma(1) - \sigma(0)
\end{aligned}$$

And the inequalities become equalities and because $\sigma(0) = 0$ we conclude $\sigma(t) = t||\mathbf{x} - \mathbf{z}||$.
Step 2 (There exists some unitary vector $\mathbf{v}$ such that $|(\partial f^*/\partial \mathbf{v})(\mathbf{x})| = 1$)
The candidate is $\mathbf{v} = \frac{\mathbf{z}-\mathbf{x}}{||\mathbf{x}-\mathbf{z}||}$, and lets compute the partial derivative

$$\begin{aligned}
\frac{\partial f^*}{\partial \mathbf{v}}(\mathbf{x}) &= \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h} \\
&= \lim_{h \to 0} \frac{\sigma(\frac{h}{||\mathbf{x}-\mathbf{z}||})}{h} = 1.
\end{aligned}$$

Then for each differentiable point x of f^* there exists an unitary vector $\mathbf{v}$ such that $|\partial f^*/\partial \mathbf{v}(\mathbf{x})| = 1$. Then by creating an orthonormal base such that $\mathbf{v}$ belongs to it we can deduce that $||\nabla f^*(\mathbf{x})|| = 1$. And this event occurs with almost surely because of Rademacher Theorem.

B.4. Proof Proposition 2

As a direct consequence of Theorem 2 we derive the next equality

$$\begin{aligned} &\inf_{\pi \in \Pi_\lambda(P_+, P_-)} \int_{\Omega\times\Omega} \left(\frac{1}{\epsilon}|\mathbf{x}-\mathbf{z}| - 2\right) d\pi + 2 \\ &= \sup_{f \in \mathrm{Lip}_{1/\epsilon}(\Omega)} \int_\Omega f(dP_+ - dP_-) - \frac{\lambda}{2}\left(\int_\Omega (1-f)_+ dP_+ + \int_\Omega (1+f)_+ dP_-\right). \end{aligned} \tag{31}$$

We denote as I the left hand side of (31) and $\Pi(P_+, P_-)$ the set of measures with marginals P_+, P_-. Now using the hypothesis (10) we derive the next inequality

$$I = \inf_{\pi \in \Pi(P_+, P_-)} \int_{\Omega\times\Omega} \left(\frac{1}{\epsilon}|\mathbf{x}-\mathbf{z}| - 2\right) d\pi + 2 = \frac{1}{\epsilon}\mathcal{W}(P_+, P_-).$$

Since $\mathrm{Lip}_{1/\epsilon}\mathcal{W}(P_+, P_-) = \sup_{f \in \mathrm{Lip}_{1/\epsilon}(\Omega)} \int_\Omega f(dP_+ - dP_-)$, we denote as $\psi_\epsilon \in \mathrm{Lip}_{1/\epsilon}(\Omega)$ the function where the supremum is achieved. Hence we derive the following inequality

$$\begin{aligned} \frac{1}{\epsilon}\mathcal{W}(P_+, P_-) &= \int_\Omega f_\lambda(dP_+ - dP_-) - \lambda\left(\int_\Omega (1-f_\lambda)_+ dP_+ + \int_\Omega (1+f_\lambda)_+ dP_-\right) \\ &\leq \int_\Omega \psi_\epsilon(dP_+ - dP_-) - \lambda\left(\int_\Omega (1-f_\lambda)_+ dP_+ + \int_\Omega (1+f_\lambda)_+ dP_-\right) \\ &= \frac{1}{\epsilon}\mathcal{W}(P_+, P_-) - \lambda\left(\int_\Omega (1-f_\lambda)_+ dP_+ + \int_\Omega (1+f_\lambda)_+ dP_-\right). \end{aligned}$$

Then $\int_\Omega (1-f_\lambda)_+ dP_+ + \int_\Omega (1+f_\lambda)_+ dP_- = 0$ and the first assert of the proof is completed. The second assertion is a straightforward consequence of the previous one.

C. Lipshitz constant for convolutional networks

C.1. Enforcing 1-Lipschitz dense layer

A neural network is a composition of linear and non-linear function. Let's study first a multilayer perceptron is defined as follows :

$$f(x) = \phi_k(W_k.(\phi_{k-1}(W_{k-1}\dots\phi_1(W_1.x))).$$

We name $L(f)$ the Lipschitz constant of a function f. As a composition of functions, the Lipschitz constant of a multilayer perceptron is upper bounded by the product of the individual Lipschitz constants:

$$L(f) \leq L(\phi_k) * L(W_k) * L(\phi_{k-1}) * L(W_{k-1}) * \dots * L(\phi_1) * L(W_1.x).$$

The most common activation functions such as ReLU or sigmoid are 1-Lipschitz. Thus, we can ensure that a perceptron is at most 1-Lipschitz by ensuring each dense layer W_k is 1-Lipschitz. Given a linear function represented by an $n \times m$ matrix W, it is commonly admitted that:

$$L(W) = ||W|| \leq ||W||_F \leq \max_{ij}(|W_{ij}|) * \sqrt{nm} \tag{32}$$

where $||W||$ is the spectral norm, and $||W||_F$ is the Frobenius norm. The initial version of WGAN [2] clips the weights of the networks. However, this is a very crude way to upper-bound the 1-Lipschitz (see equation 32). Normalizing by the Frobenius norm have also been proposed in [31]. In this paper, we use spectral normalization as proposed in [25]. At the inference step, we normalize the weights of each layer by dividing the weight by the spectral norm of the matrix:

$$W_s = \frac{W}{||W||}.$$

Even if this method is more computationally expensive than Frobenius normalization, it gives a finer upper bound of the 1-Lipschitz constraint of the layer. The spectral norm is computed by iteratively evaluating the largest singular value with the power iteration algorithm [13]. This is done during the forward step and taken into account for the gradient computation.

C.2. Enforcing 1-Lipschitz convolutional layer

In this section we will show that enforcing convolution kernels to 1-lispchitz is not enough for ensuring the 1-lipschitz property of convolutional layers, and will propose two normalization factors. Notations: We consider a Convolutional layer with an input feature map X of size (c, w, h), and L output channels obtained with kernels $W = \{W_l\}_{l \in [0,L[}$ of odd size (c, k, k), i.e. $k = 2 * \bar{k} + 1$. Considering the classical *same* configuration which output size is (L, w, h), we use the following matrix notations of the convolution $Y = W * X$:

- $\widetilde{X}$ the zero padded matrix of X of size $(c, w + k - 1, h + k - 1)$
- $\bar{W}$ the vectorized matrix of weights of size $(L, c.k^2)$
- $\bar{X}$ a matrix of size $(c.k^2, w.h)$, a duplication of the input $\widetilde{X}$, where each column j correspond to the $c.k^2$ inputs in $\widetilde{X}$ used for computing a given output j
- $\bar{Y} = \bar{W}.\bar{X}$ the vectorized output of size $(L, w.h)$

Given two outputs X_1 and X_2, we can compute an upper bound of convolutional layer lipschitz constant (Eq. 33).

$$\begin{aligned} ||Y_1 - Y_2||^2 = ||\bar{Y_1} - \bar{Y_2}||^2 &\leq ||\bar{W}||^2.||\bar{X_1} - \bar{X_2}||^2 \\ &\leq \Lambda^2.||W||^2.||X_1 - X_2||^2 \end{aligned} \tag{33}$$

The coefficient Λ^2 can be estimated, as in [8], by the maximum number of duplication of the input matrix $\widetilde{X}$ in $\bar{X}$: each input can be used at most in k^2 positions. But since within $\bar{X}$, part of the values come from the zero padded zones in $\widetilde{X}$, and have no influence on $||Y_1 - Y_2||^2$, we propose a tighter estimation of Λ, computing the average duplication factor of non zero padded value in $\bar{X}$.

For a 1D convolution (see Fig. 6), the number of zero values in the $\bar{k}$ first columns of $\bar{X}$ (symmetrically on the $\bar{k}$ last columns) is $(\bar{k}, \bar{k} - 1, ..., 1)$. So the number of zero padded values is $k.w - 2 * \sum_{t=1}^{\bar{k}} t = k.w - \bar{k}(\bar{k} + 1)$.

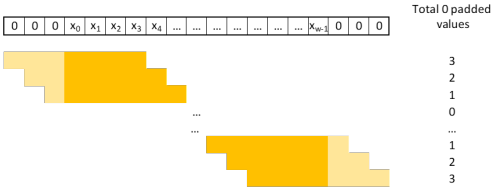

Figure 6: Zero padded elements in a 1D convolution with $k = 7(\bar{k} = 3)$

We propose to use Eq. 34 as a tighter normalization factor[2].

$$\Lambda = \sqrt{\frac{(k.w - \bar{k}.(\bar{k} + 1)).(k.h - \bar{k}.(\bar{k} + 1))}{h.w}} \tag{34}$$

C.3. Convolution layers with zero padding and stride

Convolution layers are sometimes used with stride (as in Resnet layers [17]) to reduce the computation cost of these layers[3]. Given a stride (s, s), the output layer size of the layer will be (wo, ho) such as $w = s.wo + rw$ and

[2]this factor Eq. 34 does not lead to a strict upper bound of the lipschitz constant, since particular matrix with high value on the center and low values on borders won't satisfy the inequality (33)

[3]main drawback with stride is that each point in the input feature map has not the same number of occurrences

Layer type	Parameters	Upper lip constant	Thighter Lip estimation
Dense		$\|\|W\|\|$	
Convolution wo stride	kernel size (k,k) $k=2\bar{k}+1$	$k.\|\|W\|\|$	$\sqrt{\frac{(k.w-\bar{k}.(\bar{k}+1)).(k.h-\bar{k}.(\bar{k}+1))}{h.w}}.\|\|W\|\|$
Convolution with stride	kernel size (k,k) stride (s,s)	$\lceil\frac{k}{s}\rceil.\|\|W\|\|$	$\sqrt{\frac{(k.wo-zl-zr_w).(k.ho-zl-zr_h)}{h.w}}.\|\|W\|\|$
MaxPoolig		1	
AveragePooling	averaging size po stride s	$\lceil\frac{po}{s}\rceil.\frac{1}{po}$	

Table 2: Main

$h = s.ho + rh$. We also introduce $\alpha = \lceil\frac{k}{s}\rceil$ the maximum number of overlapping stride positions. As in previous section, we can build a matrix $\bar{X}$ of size $(c.k^2, wo.ho)$, as a duplication of $\widetilde{X}$. The maximum duplication factor of an element of $\widetilde{X}$ in $\bar{X}$ is $\Lambda^2 = \alpha^2$.

As in section C.2, we can compute a tighter factor using the average duplication factor of input in X, by computing the number of non-zero-padded values used in $\bar{X}$. We introduce $\bar{\alpha}, \bar{\beta}$ such as $\bar{k} = \bar{\alpha}.s + \bar{\beta}$.

For a 1D convolution (see Fig. 7), the number of zero values in the first columns of $\bar{X}$ is $(\bar{k}, \bar{k}-s, ..., \bar{\beta})$. So the number of zero padded values on the left side is $zl = \sum_{t=0}^{\bar{\alpha}}(\bar{k} - t.s) = (\bar{\alpha}+1)\bar{k} - s.\frac{\bar{\alpha}(\bar{\alpha}+1)}{2} = \frac{(\bar{\alpha}+1)(\bar{\alpha}s+2\bar{\beta})}{2}$.

On the right side (last columns), we introduce $\gamma_w = argmax\{\gamma = w-1-i.s,$ such as $i >= 0$ and $\gamma \leq \bar{k}\}$ i.e. $\gamma_w = w-1-s.\lceil\frac{w-1-\bar{k}}{s}\rceil$. γ_w represents the first half-kernel to include the last element of the line. We also introduce α_w, β_w such as $\gamma_w = \alpha_w.s+\beta_w$. The number of zero values in the last columns is $(\bar{k}-\gamma_w, \bar{k}-\gamma_w+s, ..., \bar{k}-\gamma_w+\alpha_w.s)$, i.e. $zr_w = \sum_{t=0}^{\alpha_w}(\bar{k} - \gamma_w + t.s) = (\alpha_w+1)(\bar{k} - \gamma_w + \frac{s.\alpha_w}{2})$.

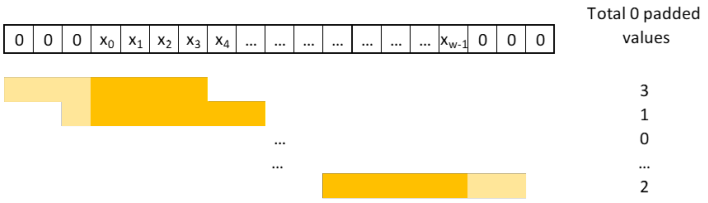

Figure 7: Zero padded elements in a 1D convolution with stride: $k = 7$ ($\bar{k} = 3$), and $s = 2$

For the matrix $\bar{Y}$ the average duplication factor for a value of the input X is $\frac{(k.wo-zl-zr_w).(k.ho-zl-zr_h)}{h.w}$
We propose to use Eq. 35 as a tighter normalization factor[4][5].

$$\Lambda = \sqrt{\frac{(k.wo - zl - zr_w).(k.ho - zl - zr_h)}{h.w}} \tag{35}$$

C.3.1 Pooling layers

By definition, the max pooling layer is 1-lipschitz, since $||max(X_1) - max(X_2)|| \leq ||X_1 - X_2||$.

Considering average pooling layer with a averaging size of po, and a stride of s. Since a mean is equivalent to a convolution with the matrix $\frac{1}{po^2}\mathbb{1}_{po\times po}$. The average pooling layer is equivalent to a convolution with stride (sec C.3). Introducing $\alpha = \lceil\frac{po}{s}\rceil$, which is 1 in the common case where $s = po$. So an upper bound of lipschitz constant for the average pooling layer is $\Lambda.||W|| = \frac{\alpha}{po}$

[4] As in previous section, this factor is not an upper bound of the lipschitz constant

[5] in case of stride $s = 1$, we have $\bar{\alpha} = \bar{k}$, $\bar{\beta} = 0$, $\gamma_w = \alpha_w = \bar{k}$ and $\beta_w = 0$. So we can retrieve $zl + zr_w = \frac{\bar{k}.(\bar{k}+1)}{2} + \frac{\bar{k}.(\bar{k}+1)}{2} = \bar{k}.(\bar{k}+1)$

C.4. Gradient norm preserving and general architecture

As proven Sections 3.2 and , the optimal function f^* with respect to Equation 5, verifies $||\nabla f^*|| = 1$ almost surely. In [16], the authors propose to add a regularization terms with respect to the average gradient norm with respect to inputs in the loss function. However, the estimation of this value is difficult and a regularization term doesn't guarantee the property. In this paper, we apply the approach described in [1], based on the use of specific activation functions and a normalization process of the weights. Three norm preserving activation functions are proposed:

- **MaxMin** : order the vector by pairs.
- **GroupSort** : order the vector by group of a fixed size.
- **FullSort** : order the vector.

These function are vector-wise rather than element-wise. We also propose the activation **ConstPReLU**, a PReLU [18] activation function complemented by a constraint such that $|\alpha| \leq 1$ (α the learnt slope). This last function is norm preserving only when $|\alpha| = 1$ (linear, or absolute value function), but being computed element wise, it is then more efficient for convolutional layers outputs.

Given a vector v of size k the P-norm pooling is defined in [4] as follows :

$$Pool_{P-norm}(v) = \left(\frac{1}{k} \sum_{i=1}^{k} v_i^p P \right)^{\frac{1}{P}}$$

Concerning gradient norm preserving linear layers, a weight matrix W is norm preserving if and only if all the singular values of W are equals to 1. In [1], the authors propose to use the Björk Orthonormalization algorithm [3]. The Björk algorithm compute the closest orthonormal matrix by repeating the following operation :

$$W_{k+1} = W_k(I + \sum_{i=1}^{p}(-1)^p \begin{pmatrix} -\frac{1}{2} \\ p \end{pmatrix} Q_k^p) \tag{36}$$

where $Q_k = 1 - W_k^T W_k$ and $W_0 = W$. This algorithm is fully differential, and as for spectral normalization, it is applied during the forward inference, and taken into account for back-propagation.

C.5. Robustness bounds

Given a 1-lipschitz neural network g and N functions compose one 1-lipschitz dense layer with a single output g_i. We consider the multi-outputs neural network $f = [g_i \circ g]_{i \in [0,N[}$, and denote $f_i = g_i \circ g$.

For a given input x of label t, we denote

$$M_f(x) = max(0, f_t(x) - max_{i \neq t}(f_i(x))$$

Theorem 4 (Adversarial Perturbation Robustness Condition under Lp Norm)**.** *If $M_f(x) > 2.\epsilon$ where $f = [g_i \circ g]_i$ is a concatenation of 1-lipschitz neural network under the L_p norm. Then x is robust to any input perturbation Δx with $||\Delta x||_p < \epsilon$*

Proof: Suppose x well classified of class t, such that $M_f(x) > 2\epsilon$. We have

$$\forall i \neq t, f_t(x) - f_i(x) \geq M_f(x) > 2\epsilon$$

Given Δx such that $||\Delta x||_p < \epsilon$, and $x' = x + \Delta x$.
Since g_i and g are 1-lipschitz, for all i, we have:

$$|\Delta y_i|^p = |g_i \circ g(x') - g_i \circ g(x)|^p \leq ||g(x') - g(x)||_p^p \leq ||\Delta x||_p^p < \epsilon^p$$

So,

$$|\Delta y_t|^p + |\Delta y_n|^p < 2.\epsilon^p$$

Layer	Number of neurons	Kernel	Output Size
Input	N/A	N/A	784x1
dense	256	N/A	256
dense	256	N/A	256
output	10	N/A	10

Table 3: MNIST dense general architecture

Layer	Number of neurons	Kernel	Output Size
Input	N/A	N/A	28x28x1
Conv	16	3x3	28x28x16
pooling	N/A	2x2	14x14x16
Conv	32	3x3	14x14x32
pooling	N/A	2x2	7x7x32
dense	100	N/A	100
output	10	N/A	10

Table 4: MNIST CNN general architecture

$$g(\frac{|\Delta y_n| + |\Delta y_t|}{2}) \leq g(|\Delta y_t|) \iff \frac{(|\Delta y_n| + \Delta y_t|)^p}{2^{p-1}} \leq |\Delta y_t|^p + |\Delta y_n|^p < 2.\epsilon^p$$

So, $\forall n \neq t$

$$y'_t - y'_n = y_t - y_n + \Delta y_t - \Delta y_n \geq M_f(x) - (|\Delta y_n| + |\Delta y_t|) > M_f(x) - 2\epsilon > 0$$

So for all Δx such that $||\Delta x||_p < \epsilon$, and $x' = x + \Delta x$, x' is classified as t.

In [22], authors report a provable robustness of $\sqrt{2}$. However, in their case the global classifier with the N outputs is 1-lipschitz meaning that the f_i have a lipschitz constant lesser than 1. Then, in their case the maximal value of $M_f(x)$ is lesser than the one of our network, making the comparison of the robustness provable constants not possible directly.

D. Experiments : additional results

D.1. Networks architecture

In order to have a fair comparison of the competitors, we use the same architectures for the neural network given a dataset. The architectures are described in Tables 3, 4, 5 and 6. The activation functions, pooling functions, and normalization functions are described in Tables 7, the optimizations parameters in Table 8 and the attacks parameters in Table 9.

Layer	Number of neurons	Kernel	Output Size
Input	N/A	N/A	32x32x3
Conv	128	3x3	32x32x128
Conv	128	3x3	32x32x128
pooling	N/A	2x2	16x16x128
Conv	256	3x3	16x16x256
Conv	256	3x3	16x16x256
pooling	N/A	2x2	8x8x256
Conv	512	3x3	8x8x512
Conv	512	3x3	8x8x512
pooling	N/A	2x2	4x4x512
dense	512	N/A	512
dense	512	N/A	512
output	10	N/A	10

Table 5: CIFAR CNN general architecture

Layer	Number of neurons	Kernel	Output Size
Input	N/A	N/A	128x128x3
Conv	16	3x3	128x128x16
Conv	16	3x3	128x128x16
Conv	16	3x3	128x128x16
pooling	N/A	2x2	64x64x16
Conv	32	3x3	64x64x32
Conv	32	3x3	64x64x32
Conv	32	3x3	64x64x32
pooling	N/A	2x2	16x16x32
Conv	64	3x3	16x16x64
Conv	64	3x3	16x16x64
Conv	64	3x3	16x16x64
pooling	N/A	2x2	8x8x64
Conv	128	3x3	8x8x128
Conv	128	3x3	8x8x128
Conv	128	3x3	8x8x128
pooling	N/A	2x2	4x4x128
Conv	256	3x3	4x4x256
Conv	256	3x3	4x4x256
Conv	256	3x3	4x4x256
pooling	N/A	2x2	2x2x256
dense	512	N/A	512
dense	512	N/A	512
output	10	N/A	10

Table 6: CelebA CNN general architecture

Network	**Conv activation**	**Dense activation**	**Output activation**	**Pooling**	**Orthonormalization**
Adv	ReLU	ReLU	softmax	Maxpooling	None
$1LIP$	ReLU	ReLU	softmax	Maxpooling	Björck
GNP_{log}	GroupSort2	Fullsort	softmax	2-norm	Björck
GNP_{hin}	GroupSort2	Fullsort	linear	2-norm	Björck
hKR	GroupSort2	Fullsort	linear	2-norm	Björck

Table 7: Algorithms specific features

Dataset	**Optimizer**	**Steps per epoch**	**Nb epochs**	**Learning rate**	**Batch size**	**Augmentation**
MNIST dense	Adam	60000	100	0.01	256	no
MNISTY conv	Adam	60000	100	0.01	256	no
CIFAR 10	Adam	45000	100	.00001	256	no
CELEB A	Adam	10000	200	0.0005	64	yes

Table 8: Optimization parameters

Dataset	l_2 **deepfool**	l_2 **FGM**	l_2 **PGD**	l_2 **CW**
MNIST dense	$\epsilon \in \mathbb{R}^+$ 2000 attacks	ϵ from 0.1 to 7.9 st. 0.1 500 attacks	ϵ from 0.1 to 7.9 st. 0.1 500 attacks	$\epsilon \in [0, 1, 2, 4, 6, 8]$ 500 attacks
MNIST conv	$\epsilon \in \mathbb{R}^+$ 2000 attacks	ϵ from 0.1 to 7.9 st. 0.1 500 attacks	ϵ from 0.1 to 7.9 st. 0.1 500 attacks	$\epsilon \in [0, 1, 2, 4, 6, 8]$ 500 attacks
CIFAR 10	$\epsilon \in \mathbb{R}^+$ 2000 attacks	ϵ from 0.1 to 8 st. 0.2 500 attacks	ϵ from 0.1 to 8 st. 0.2 500 attacks	$\epsilon \in [0, 1, 2, 4, 6, 8]$ 500 attacks
CELEB A	$\epsilon \in \mathbb{R}^+$ 2000 attacks	$\epsilon \in [2, 5, 7]$ 500 attacks	$\epsilon \in [2, 5, 7]$ 500 attacks	$\epsilon \in [2, 5, 7]$ 500 attacks

Table 9: ϵ values for the different dataset and attacks